Stereo Matching Using Epipolar Distance Transform

Qingxiong Yang, *Member, IEEE*, and Narendra Ahuja, *Fellow, IEEE*

***Abstract*— In this paper, we propose a simple but effective image transform, called the epipolar distance transform, for matching low-texture regions. It converts image intensity values to a relative location inside a planar segment along the epipolar line, such that pixels in the low-texture regions become distinguishable. We theoretically prove that the transform is affine invariant, thus the transformed images can be directly used for stereo matching. Any existing stereo algorithms can be directly used with the transformed images to improve reconstruction accuracy for low-texture regions. Results on real indoor and outdoor images demonstrate the effectiveness of the proposed transform for matching low-texture regions, keypoint detection, and description for low-texture scenes. Our experimental results on Middlebury images also demonstrate the robustness of our transform for highly textured scenes. The proposed transform has a great advantage, its low computational complexity. It was tested on a MacBook Air laptop computer with a 1.8 GHz Core i7 processor, with a speed of about 9 frames per second for a video graphics array-sized image.**

***Index Terms*— Epipolar, stereo matching, texture.**

I. Introduction

COMPUTATIONAL stereo for extraction of three-dimensional scene structure has traditionally been, and continues to be an active area of intense research interest [1], [2]. In the past decade, much of the community's effort has been focused on the specific problem of *disparity optimization*, producing a number of excellent optimization methods that have significantly advanced the state of the art. The key objective of these optimization methods is to reduce the matching ambiguities introduced by low-texture regions, and they can be generally classified into three categories: local methods, global methods, and hybrid methods.

The best *local* methods known today are either based on edge-preserving filtering or image segmentation. Yoon and Kweon [3] aggregate the matching cost with respect to both the color similarity and geometric proximity, and Yang [4] is the first to propose a non-local cost aggregation algorithm based on a minimum spanning tree computed from the reference camera image. Zitnick *et. al.* [5] aggregate the matching cost within each image segment and the obtained disparity maps from different cameras located at different positions are then fused to give coherent estimates.

The most popular *global* methods are based on belief propagation [6], [7] or graph cuts [8]. Both methods are formulated in an energy-minimization framework [9], where the objective is to find a disparity solution that minimizes a global energy function.

In low-texture regions, the lack of visual features makes matching a challenging problem. Local methods, which typically assume that the disparity values are the same for pixels inside the support window, do not work well on non-fronto-parallel surfaces which do not satisfy this assumption. Assuming that only the neighboring pixels have the same disparity value, global methods are more suitable for non-fronto-parallel surfaces, but only when the size of the low-texture regions is relatively small. Several hybrid methods [10]–[13] have been proposed to take advantage of both local and global optimization techniques. These methods assume planar surfaces and alternate between assigning pixels to 3D planes and refining the plane equations. A common problem with these techniques is that they rely on having accurate image segmentation, which may not always be robust. Other optimization methods make more restricted assumptions like Manhattan-world [14], [15], or only extracts vertical facades [16], [17].

All of the above methods greatly advance the state of stereo vision in the indicated ways, they are adversely affected by the noisy nature of the matching cost as computed from the image intensities while neglecting the image structure as a source of obtaining robust matching invariants.

In this paper, we propose an image transform - epipolar distance transform - which helps estimate planar 3D structure at points in low-texture areas in terms of distances measured along the epipolar lines. Specifically, we extract the boundaries of a homogeneous (low-texture) region and locate its two points of intersection with the epipolar line. We next compute the distance between the two endpoints and the distance between one endpoint and every pixel on the within-region epipolar line segment. For planar surfaces, the ratio of the distances is invariant to affine transformation, and thus can be used as a matching invariant for stereo vision.

Unlike image intensity/color, our transform is robust for matching low-texture regions. Note that our transform is proposed to improve the robustness of existing stereo algorithms in matching low-texture regions; we are not claiming a new/better stereo algorithm. Our experiments with both global stereo method (belief propagation: CSBP [18]) and local stereo method (sliding window) demonstrate the effectiveness of our transform.

Manuscript received July 11, 2011; revised May 22, 2012; accepted June 18, 2012. Date of publication July 10, 2012; date of current version September 13, 2012. This work was supported in part by Hewlett-Packard under the Open Innovation Research Program and a startup grant from the City University of Hong Kong under Project 7200250. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Stefan Winkler.

Q. Yang is with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: liiton.research@gmail.com).

N. Ahuja is with the Department of Electrical and Computer Engineering, Becdkman Institute and Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61820 USA (e-mail: n-ahuja@illinois.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2012.2207393

1057–7149/$31.00 © 2012 IEEE

A. Assumption and Justification

The epipolar distance transform proposed in this paper is derived based on affine transforms. Thus it is theoretically valid only for planar surfaces, an assumption actually used in almost every state-of-the-art stereo algorithm [10]–[16]. *This assumption is harmless in textured parts of an image since texels are, by definition, small and many compared to image size* (see Fig. 9), and it is safe to assume that small regions can be well-approximated as corresponding to planar surfaces. The higher the curvature, i.e., the degree of violation of planarity assumption, by a curved surface, low-texture region, the higher will be the resulting reconstruction error. For Lambertian surfaces, higher curvature also means greater variation due to shading (larger variance of surface normals). We can use the extent of this intensity variation to select virtual endpoints (of a virtual line segment). The endpoints are virtual in that they are not true ends of a line segment corresponding to a planar region; rather, the intensity variation within the line segment is acceptably close to that which would be found if the surface was indeed planar. We can estimate the probability that two given points are virtual endpoints from the intensity/color variation across the segment (see Eqn. 4). The larger the curvature, the more confident will be the probability estimate. The virtual line segments will partition the whole curved surface into a number of small, low-curvature patches, corresponding to a polyhedral approximation. The details are presented in Sec. II. The experimental results in Fig. 6 demonstrate that due to such automatic polyhedralization of a curved surface, the proposed epipolar distance transform performs robustly for non-planar, low-texture surfaces, although we do use planarity to derive the basic theory.

II. Approach

Lack of texture leads to ambiguities/errors in matching when the image intensity/color is used as the matching invariant. On the other hand, the geometric properties of the image segments, such as area, boundary shape, and mutual distances, are more robust to the intensity variations. Ahuja [19] proposed an image transform to capture multiscale image structure by computing an attraction-force field over the image. This transform can be used to extract a multi-scale segmentation tree, which has been proven to be very efficient for object categorization [20], [21]. However, the force values constituting the image transform reflect strictly 2D structure. Not being affine invariant, the force values at points within low texture regions cannot be used as features for stereo matching.

A. Epipolar Distance Transform

In this section, we present a new image transform - epipolar distance transform - to capture the image structure. This transform is invariant to affine transform, and can be used as a matching invariant for stereo vision. To define the transform, let **PQ** in Fig. 1(a) be a straight line segment in Euclidean 3-space $\mathbb{R}^3$, and **O** be a point inside **PQ**, then their projections on the left camera C_L and right camera C_R have the following property

$$\frac{\| \mathbf{P}_L - \mathbf{O}_L \|}{\| \mathbf{P}_L - \mathbf{Q}_L \|} = \frac{\| \mathbf{P}_R - \mathbf{O}_R \|}{\| \mathbf{P}_R - \mathbf{Q}_R \|} \tag{1}$$

where $\mathbf{P}_L$, $\mathbf{O}_L$ and $\mathbf{Q}_L$ are the projections of **P**, **O** and **Q** on camera C_L, and $\mathbf{P}_R$, $\mathbf{O}_R$ and $\mathbf{Q}_R$ are projects on camera C_R. We can establish Eqn. (1) because the ratios of lengths are preserved under affine transform, and the mapping between line $\mathbf{P}_L\mathbf{Q}_L$ and $\mathbf{P}_R\mathbf{Q}_R$ is an affine transform as long as **PQ** is a straight line segment according to the epipolar geometry.

To make the problem simpler, we can assume that the camera motion is pure translation, or equally the camera images are rectified such that the epipolar lines are scanlines, and $\mathbf{P}_L\mathbf{Q}_L$ and $\mathbf{P}_R\mathbf{Q}_R$ are line segments along the epipolar line on camera C_L and C_R, respectively. To compute the ratio of lengths at each pixel location $\mathbf{O}_L$, we need to detect two endpoints $\mathbf{P}_L$ and $\mathbf{Q}_L$ of the line segment $\mathbf{P}_L\mathbf{Q}_L$ passing through $\mathbf{O}_L$ to measure the lengths $\| \mathbf{P}_L - \mathbf{O}_L \|$ and $\| \mathbf{P}_L - \mathbf{Q}_L \|$. In theory, we can segment the image into regions, then compute the lengths as the sum of the pixels on each region along each scanline. Let scanline width be $\mathbf{w}$, x be x-axis value of a pixel $\mathbf{x}$ along scanline, and $x_L^{\mathbf{O}}$ be x-axis value of pixel $\mathbf{O}_L$,

$$\| \mathbf{P}_L - \mathbf{O}_L \| = \sum_{x=0}^{x_L^{\mathbf{O}}} step(x, x_L^{\mathbf{O}}) \tag{2}$$

$$\| \mathbf{P}_L - \mathbf{Q}_L \| = \sum_{x=0}^{\mathbf{w}-1} step(x, x_L^{\mathbf{O}}) \tag{3}$$

where

$$\text{step } (x, x_L^{\mathbf{O}}) = \begin{cases} 1 & \mathbf{x} \text{ and } \mathbf{O}_L \text{ are in the same segment,} \\ 0 & \text{otherwise.} \end{cases}$$

Fig. 2 presents a synthetic scene containing only white and black pixels, and accurate segmentation is guaranteed. P_L and Q_L are the intersections of the epipolar line passing pixel O_L and the boundary of the white regions in Fig. 2 (a). The length of line segment P_LQ_L (or P_LO_L) is equal to the number of white pixels between P_L and Q_L (or P_L and O_L). Thus for every pixel along on the same epipolar line, we need to check whether it is a white pixel or not, and this is formulated as the summation of a step function in Eqn. (3). The ratio of the lengths computed using Eqn. (1) is presented in Fig. 2 (c)–(d). However, it is impossible to separated a real image into only white and black pixels, thus a $step()$ function is not practical.

Besides, image segmentation is known to be non-robust. For instance, the detected line segment $\mathbf{P}_L\mathbf{Q}_L$ on the left image may be split into multiple line segments in the right image. Such an example is shown in Fig. 3. The EDISON system (Meanshift segmentation) [22] with default parameters is used to segment the left and right images in Fig. 6(a) and (b), and the results are presented in Fig. 3. The white boxes indicate where the system obtains inconsistent segmentation results on the left and right images.

In practice, we adopt a "soft segmentation" approach. Specifically, we replace the $step()$ function in Eqn. (2) and (3)

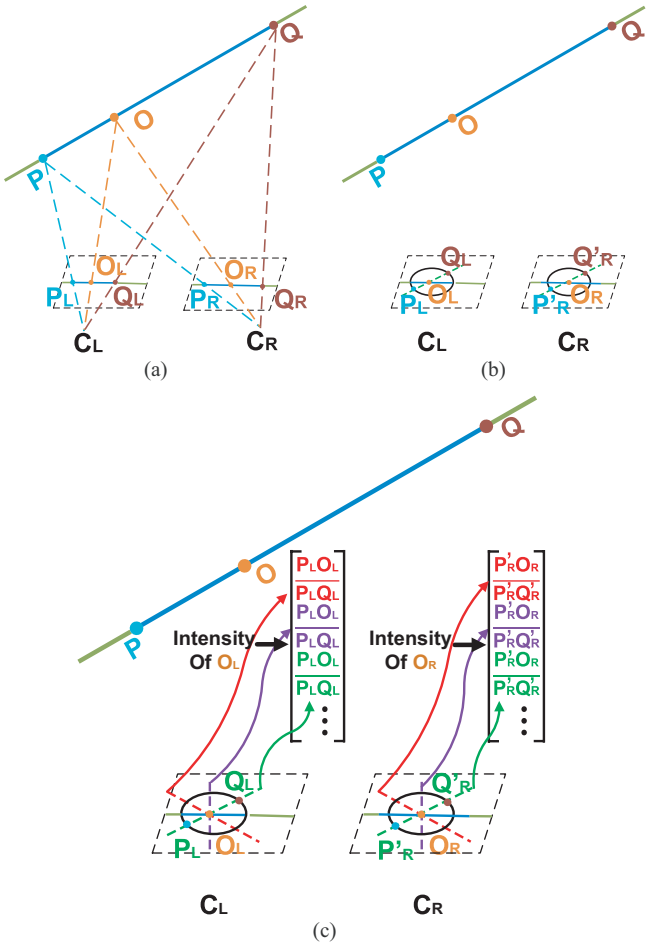


Fig. 1. (a) **PQ** is a line segment in $\mathbb{R}^3$, and its projections on camera C_L and C_R are $\mathbf{P}_L\mathbf{Q}_L$ and $\mathbf{P}_R\mathbf{Q}_R$. The mapping between $\mathbf{P}_L\mathbf{Q}_L$ and $\mathbf{P}_R\mathbf{Q}_R$ is an affine invariant. (b) Projections of a line segment **PQ** passing **O** can be detected by drawing a straight line passing through projection of **O**, and the intersections of this line and the region boundary are projections of **P** and **Q**. (c) Projections of different line segments passing **O** can be detected by drawing straight lines passing through projection of **O** in different directions. Please zoom in to see the details if it is unclear.

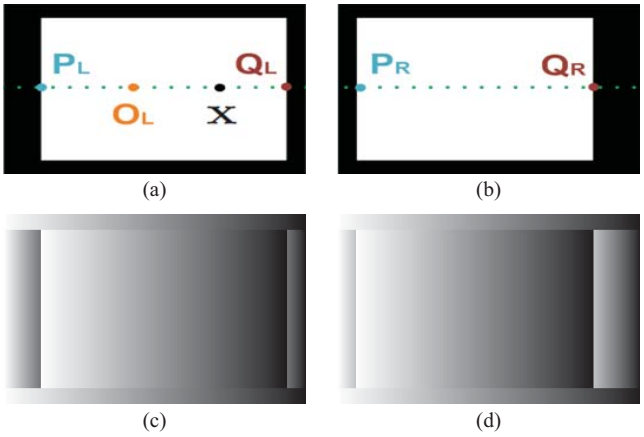


Fig. 2. Synthetic scene. (a) and (b) Left and right images and (c) and (d) transformed images.

with a Gaussian function with standard deviation σ_I:

$$g(I(x), I(x_L^{\mathbf{O}})) = exp\left(-\frac{(I(x) - I(x_L^{\mathbf{O}}))^2}{2\sigma_I^2}\right) \quad (4)$$

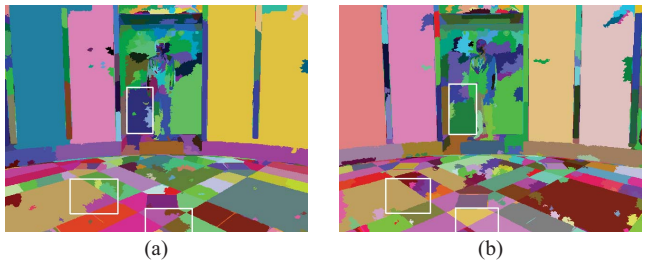


Fig. 3. Meanshift segmentation. The white boxes indicate where the system obtains inconsistent segmentation results on the left and right images. (a) Left. (b) Right.

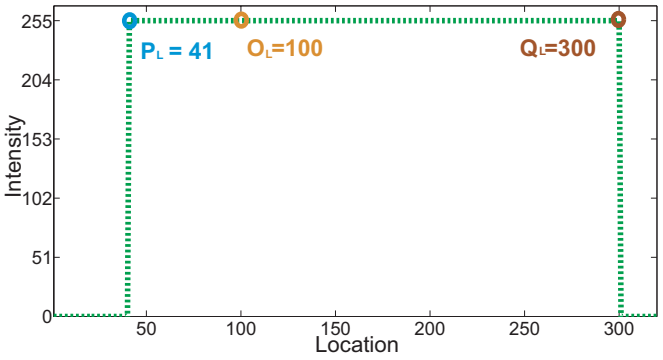


Fig. 4. Scanline of Fig. 2(a).

where $I(x)$ and $I(x_L^{\mathbf{O}})$ are the intensity values of pixel $\mathbf{x}$ and $\mathbf{O}_L$. $g(I(x), I(x_L^{\mathbf{O}}))$ increases as the intensities of $\mathbf{x}$ and $\mathbf{O}_L$ get closer. Unlike the $step()$ function in Eqn. (2) and (3) which gives a binary decision, $g(I(x), I(x_L^{\mathbf{O}}))$ measures the similarity of pixel $\mathbf{x}$ and O_L based on their intensity similarity, and gives a values between zero and one. However, $g(I(x), I(x_L^{\mathbf{O}}))$ will be the same as the $step()$ function if $\sigma_I = 0$. σ_I should be small enough to suppress contributions of dissimilar pixels (from different regions). But to account for sensor noise, we set σ_I to 7 for all 8-bit images used in our experiments. Eqn. (4) is an approximation of the step() function, but more robust for real images.

Unlike standard methods, we do not require the intensities of the correspondences to be the same. Let x_R and $x_R^{\mathbf{O}}$ be the x-axis values of the correct correspondences of pixel $\mathbf{x}$ and $\mathbf{O}_L$ in camera C_R, respectively. As can be seen from Eqn. (2), Eqn. (3) and Eqn. (4), we only require $g(I(x_R), I(x_R^{\mathbf{O}})) = g(I(x), I(x_L^{\mathbf{O}}))$, that is the intensity difference of x_R and $x_R^{\mathbf{O}}$ should be the same as the intensity difference of x and $x_L^{\mathbf{O}}$ to make sure that the ratio computed from the summations in Eqn. (2) and Eqn. (3) is affine invariant. As a result, our method is more robust to brightness changes. The existence of noise in real images violates this assumption, but our experiments show that this technique is robust to noise when σ_I is sufficiently large, e.g., $\sigma_I = 7$ (for 8-bit images) in our experiments.

The discussion above shows that the matching invariant computed at each pixel $\mathbf{O}_L$ is actually an intensity-weighted summation of ones along the epipolar line (substitute Eqn. 4

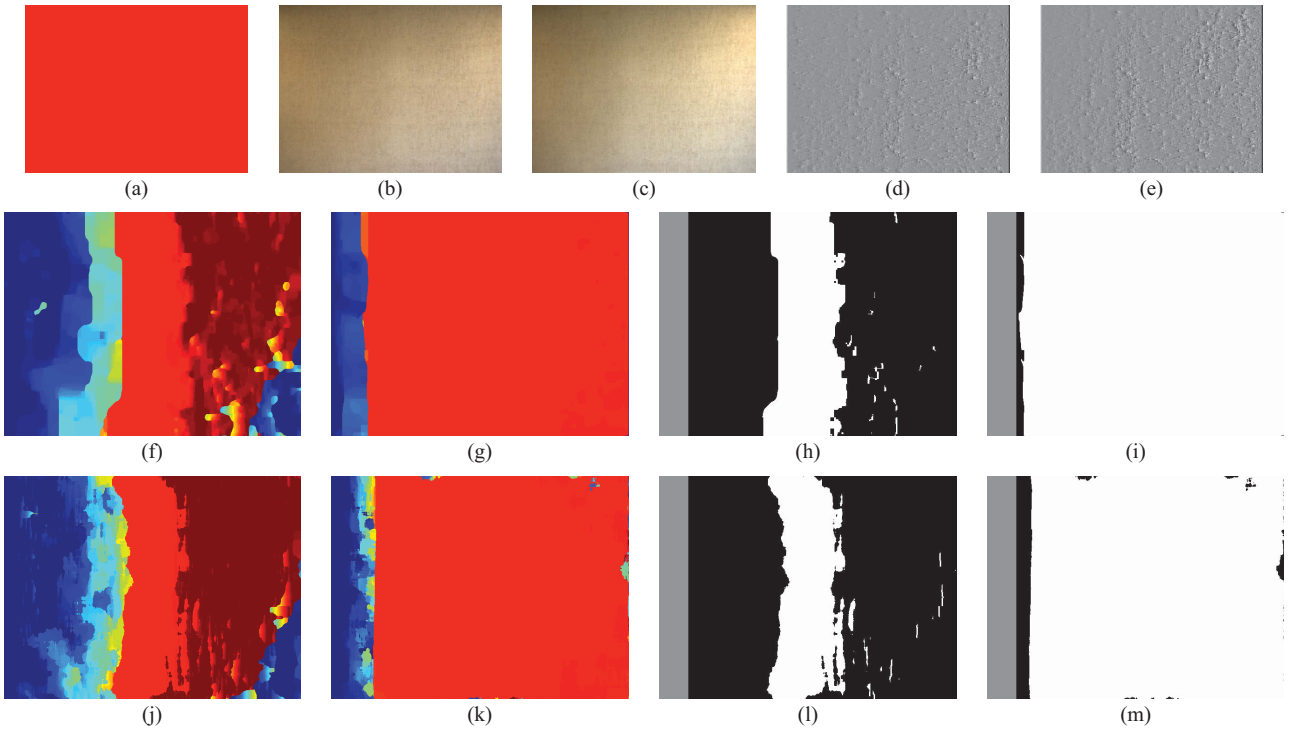


Fig. 5. Fronto-parallel surfaces. (a) Measured disparity map, (b) and (c) input stereo images, and (d) and (e) images after transform. (f) and (g) Disparity maps obtained by applying global stereo method (CSBP [18]) to (b) and (c), and (d) and (e), respectively. (h) and (i) Disparity error maps of (f) and (g), respectively. The black pixels are bad pixels with an error larger than one disparity. The gray pixels are border-occluded, and are not considered in this paper. Apparently, our transform performs much better than image intensity for matching low-texture wall surfaces using both global and local stereo methods. (j)–(m) Results obtained using local sliding window stereo method. The quantitative comparison is summarized in Table I.

into Eqn. 2 and 3):

$$\mathcal{F}(\mathbf{O}_L) = \frac{\| \mathbf{P}_L - \mathbf{O}_L \|}{\| \mathbf{P}_L - \mathbf{Q}_L \|} = \frac{\sum_{x=\max(0,x_L^{\mathbf{O}}-\sigma_S \mathbf{w})}^{x_L^{\mathbf{O}}} e^{-\frac{(I(x)-I(x_L^{\mathbf{O}}))^2}{2\sigma_I^2}} \cdot 1}{\sum_{x=\max(0,x_L^{\mathbf{O}}-\sigma_S \mathbf{w})}^{\min(\mathbf{w}-1,x_L^{\mathbf{O}}+\sigma_S \mathbf{w})} e^{-\frac{(I(x)-I(x_L^{\mathbf{O}}))^2}{2\sigma_I^2}} \cdot 1} \quad (5)$$

where σ_S is a scalar controlling the size of the local regions to be taken into account. Note that Eq. (5) is similar to a joint bilateral filter kernel and maybe further accelerated using the techniques proposed in [23], [24].

An example of computing the matching invariant using Eq. (5) is presented in Fig. 4. The green dash line in Fig. 4 is a scanline extracted from Fig. 2(a). The horizontal axis corresponds to the pixel locations (0 to 319), and the vertical axis corresponds to the intensity values. As can be seen in Fig. 4, $\| \mathbf{P}_L - \mathbf{O}_L \| = 60$ and $\| \mathbf{P}_L - \mathbf{Q}_L \| = 260$, thus the ratio $\frac{\|\mathbf{P}_L-\mathbf{O}_L\|}{\|\mathbf{P}_L-\mathbf{Q}_L\|} = \frac{60}{260}$ is the ground-truth matching invariant at pixel $\mathbf{O}_L$ according to Eq. (1). Let $\sigma_s = +\infty$ and $\sigma_I = 7$, the matching invariant can be approximated using Eq. (5):

$$\mathcal{F}(\mathbf{O}_L) = \left(\sum_{x=0}^{40} e^{-\frac{(0-255)^2}{2\cdot 7^2}} + \sum_{x=41}^{100} e^{-\frac{(255-255)^2}{2\cdot 7^2}}\right) \Big/ \left(\sum_{x=0}^{40} e^{-\frac{(0-255)^2}{2\cdot 7^2}} + \sum_{x=41}^{300} e^{-\frac{(255-255)^2}{2\cdot 7^2}} + \sum_{x=301}^{319} e^{-\frac{(0-255)^2}{2\cdot 7^2}}\right) = \frac{41 \cdot e^{-663.52} + 60}{60 \cdot e^{-663.52} + 260} \sim \frac{60}{260}. \quad (6)$$

TABLE I

QUANTITATIVE EVALUATION USING STEREO IMAGES PRESENTED IN FIGS. 5(b) AND (c) AND 6(b) AND (c)

Data set	Method			
	Local stereo		CSBP [18]	
	Intensity	Ours	Intensity	Ours
Fig. 5(b)–(c)	68.1	**3.02**	68.1	**2.39**
Fig. 6(b)–(c)	37.5	**10.56**	34.2	**3.76**
Wall surface in Fig. 6	24.9	**2.56**	26.6	**2.37**

The numbers are the percentage of pixels with misestimated disparities. The second and third columns are results from the local sliding window stereo method and the fourth and fifth columns are results from the CSBP [18] method. The second and fourth columns are results computed using image intensity and the third and fifth columns are results computed using our transform. The last row contains results for only the cylindrical wall surface in Fig. 6(b). Our transform greatly reduces the percentage of bad pixels due to lack of texture. The local stereo method results in a much larger percentage of bad pixels (10.56%) using the data set presented in Fig. 6(b) and (c). This is because local stereo method assumes fronto-parallel surfaces which is invalid for this data set. Note that there is a floor surface in Fig. 6(b) and (c).

Note that the computation of $\mathcal{F}(\mathbf{O}_L)$ does not require the detection of the exact endpoints of the line segment (the location of pixel $\mathbf{P}_L$ and $\mathbf{Q}_L$).

In this paper, we set σ_S to a constant value (0.01) for real images, and use Eqn. (5) to compute the matching invariant $\mathcal{F}$ at each pixel location, which is essentially an image transform. Our experiments (Sec. III) demonstrate that stereo matching using the transformed images with constant σ_S can perform much better than standard intensity-based stereo matching methods for low-texture regions. See Sec. III for details. Note that in this case, the length of the line segment is assumed to

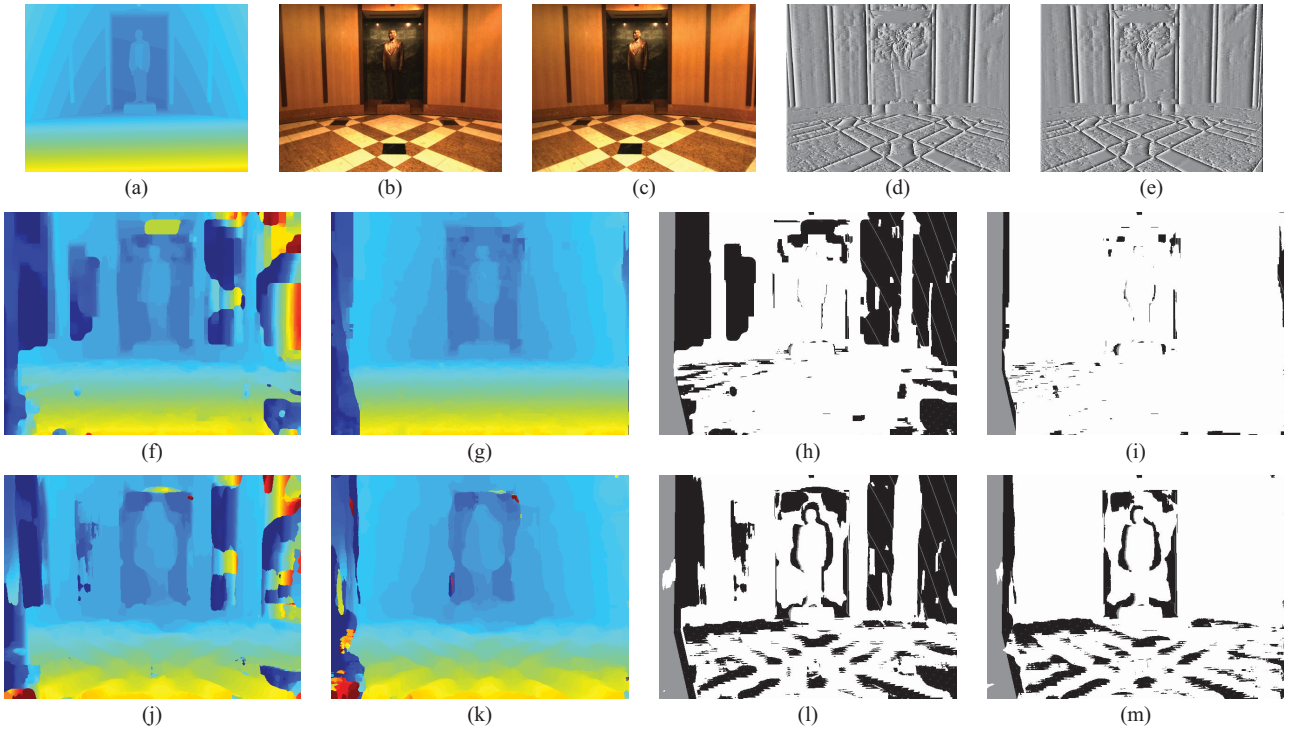


Fig. 6. Cylindrical surfaces. (a) Measured disparity map, (b) and (c) input stereo images, and (d) and (e) images after transform. (f) and (g) Disparity maps obtained by applying global stereo method (CSBP [18]) to (b) and (c) and (d) and (e), respectively. (h) and (i) Disparity error maps of (f) and (g), respectively. The black pixels are bad pixels with an error larger than one disparity. The gray pixels are border-occluded, and are not considered in this paper. Apparently, our transform performs much better than image intensity for matching low-texture wall surfaces using both global and local stereo methods. (j)–(m) Results obtained using local sliding window stereo method. The quantitative comparison is summarized in Table I.

be always less than $2 \times \sigma_S \mathbf{w} + 1$, and the detection of the exact endpoints of the line segment is avoid.

B. Other Transforms

In this section, we discuss the possibility of using the ratio of lengths along directions other than the epipolar lines. According to Eqn. (1), to compute the ratio of lengths at a pixel location $\mathbf{O}_L$ on camera C_L, we need to detect the two endpoints of a straight line passing through $\mathbf{O}_L$. This can be done using image segmentation. Assume that after segmentation, every pixel inside the black circle in Fig. 1(b) belongs to the same segment. Then we simply need to draw a straight line [*e.g.*, the green dashed line in Fig. 1(b)] passing through $\mathbf{O}_L$. The intersections of this line and the segment boundary are at $\mathbf{P}_L$ and $\mathbf{Q}_L$. We can draw lines in different directions as shown in Fig. 1(c) [*e.g.*, the red, purple and green dashed lines in Fig. 1(c)], and then compute a ratio of lengths (Eqn. 1) for every direction. For each pixel $\mathbf{O}_L$, we thus convert the intensity/color to a vector comprising of the ratios of lengths computed from different directions as depicted in Fig. 1(c).

However, a direction in the left image may not always correspond to the same direction in the right image as can be seen from Theorem 1 (the proof is provided in **Appendix**). That is the detected line segment $\mathbf{P}'_R\mathbf{Q}'_R$ (computed using the same direction as $\mathbf{P}_L\mathbf{Q}_L$) in C_R in Fig. 1(c) may not be always the correct correspondence of $\mathbf{P}_L\mathbf{Q}_L$. In this case, the ratio of lengths computed from $\mathbf{P}'_R\mathbf{Q}'_R$ won't be the same as the ratio of lengths computed from $\mathbf{P}_L\mathbf{Q}_L$, and cannot be used as a matching invariant.

Theorem 1: Let $\mathbf{PQ}$ be a straight line segment in Euclidean 3-space $\mathbb{R}^3$, $\mathbf{P}_L\mathbf{Q}_L$ be its projection on left camera C_L, and $\mathbf{P}_R\mathbf{Q}_R$ be its projection on right camera C_R, then line $\mathbf{P}_L\mathbf{Q}_L$ has the same direction as $\mathbf{P}_R\mathbf{Q}_R$ only when

1) $\mathbf{P}_L\mathbf{Q}_L$ and $\mathbf{P}_R\mathbf{Q}_R$ are along the epipolar line,
2) or the disparity values of pixel $\mathbf{P}_L$ and $\mathbf{Q}_L$ are the same.

Theorem 1 shows that for directions other than the epipolar line, we require fronto-parallel surfaces which depends on the structure of the scene to be captured and the camera orientation. The epipolar line direction, however, is independent of the scene structure and the camera orientation, thus is adopted in this paper. That is, in practice, we only use ratios of lengths computed from a single direction: along epipolar line.

III. Experiments

In this section, we present experiments on real images to demonstrate the effectiveness and robustness of our method. Sec. III-A numerically evaluate the performance of our image transform for stereo matching using two real indoor data sets, Sec. III-B presents visual evaluation on three outdoor data sets, and Sec. III-C shows that our image transform is robust to keypoint detection and description for low-texture scenes. These images are captured by a commercial stereo vision system: Point Gray Bumblebee XB3 stereo vision system [25]. Similar to the other systems, the lenses of Bumblebee XB3

stereo vision system exhibit optical vignetting to some degree, which cause problems for stereo matching especially for low-texture regions. Finally, we show that our transform is also robust for highly-textured scenes in Sec. III-D.

All the experiments conducted use the same parameters. Specifically, we set σ_I to 7 and σ_S to 0.01, and use Eqn. 5 to compute the transform. All the disparity maps presented in this section are computed using either local sliding window stereo method or Yang's CSBP method [18].[1] CSBP is a very efficient belief propagation algorithm with memory cost invariant to the disparity search range, and is employed because our laptop computer cannot afford the huge memory cost required by standard BP algorithm for the high resolution stereo images used in our experiments.

A. Numerical Evaluation Using Indoor Scenes

In this section, we numerically evaluate the performance of our image transform for stereo matching using two real indoor data sets with regions that are weakly textured. We first evaluate our method with a low-texture wall as shown in Fig. 5(b)–(c). We adjusted the camera to make sure that the wall is fronto-parallel such that the z-depth values are the same for every pixel and can be manually measured. The measured disparity map is presented in Fig. 5(a). Fig. 5(d) and (e) are the transformed images of (b) and (c). Fig. 5(f) and (g) are the disparity maps obtained by applying global stereo method (CSBP [18]) to (b)–(c) and (d)–(e), respectively, and the corresponding disparity error maps are presented in (h) and (i). As can be seen in Fig. 5(h), most of the pixels are **black** which correspond to pixels with disparity error lager than 1 (**gray** pixels are border-occluded.). These errors are corrected using our transform as shown in Fig. 5(i) except for the border occlusion (gray pixels) on the left which is not considered in this paper. (j)–(m) are the results obtained using local sliding window stereo method, and the quantitative comparison is summarized in Table I.

Fig. 6 presents the reconstruction results for a cylinder-like lobby. The camera was placed in the center of the lobby and the angle of the camera was adjusted such that the z-depth values of the pixels in each column are the same and thus the cylindrical wall surface can be measured. We then manually segmented the reference image in Fig. 6(b), and applied plane fitting to each segment using the keypoints detected and matched using SIFT [26] **except for the cylindrical wall surfaces**. The obtained disparity map is presented in Fig. 6(a). The manual image segmentation result will not be used to compute the proposed image transform. The disparity error maps presented in Fig. 6(h)–(i) and (l)–(m) show that most of the errors (**black** pixels) due to lack of texture are corrected using our transform. The numerical comparison is presented in Table I which shows that our image transform greatly improve the reconstruction accuracy. The disparity values of the **non-cylindrical** surfaces in Fig. 6(a) are estimated using plane fitting, and thus not very precise. However, most of reconstruction errors reside in the low-texture regions inside the cylindrical wall surface, we thus also limited our evaluation on only the cylindrical wall surface where the accuracy is guaranteed and present the results in the last row of Table I.

[1]We use the source code published on the author's website.

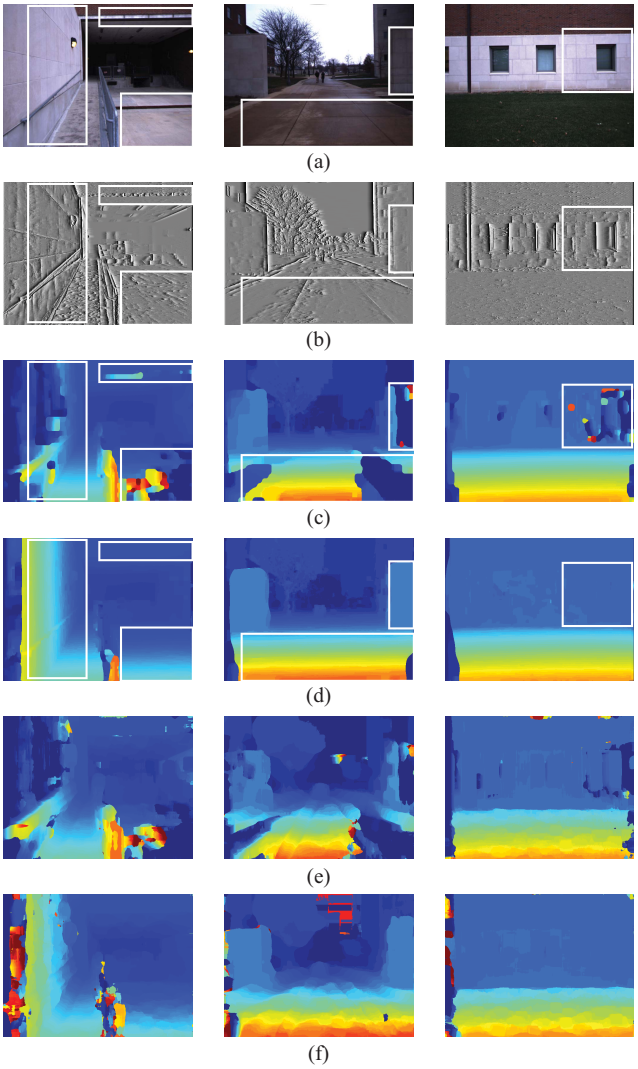


Fig. 7. Visual evaluation using outdoor scenes. (a) Reference images. (b) Transformed images. (c) Global stereo disparity maps obtained from (a). (d) Global stereo disparity maps obtained from (b). (e) Local stereo disparity maps obtained from (a). (f) Local stereo disparity maps obtained from (b). The white boxes indicate where intensity matching method fails due to lack of texture.

Note that the wall is not a planar surface, which violates the assumption we made for deriving the basic theory. However, as discussed in Sec. I-A, our transform computed from the intensity-weighted summations (Eqn. 5) is robust to curved surfaces. The disparity map in Fig. 6(g) experimentally verify this claim.

B. Visual Evaluation Using Outdoor Scenes

Fig. 7 presents the experimental results on three outdoor scenes. From top to bottom are the reference images (a), transformed images (b), disparity maps obtained by applying CSBP [18] to the input image pairs (c) and the transformed images (d), and disparity maps obtained by applying local sliding window stereo method to the input image pairs (e) and the transformed images (f), respectively. The ground-

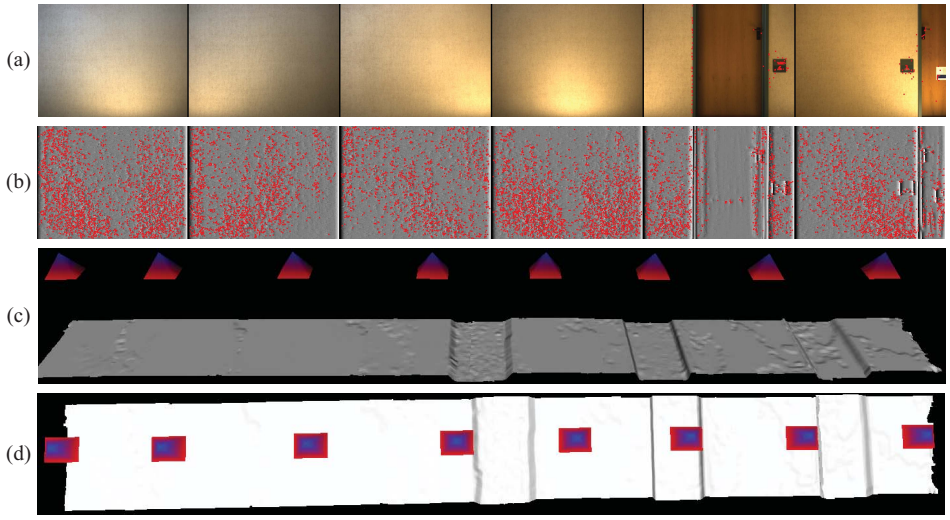


Fig. 8. Camera tracking. (a) Selected frames of a low-texture wall, (b) transformed images, and (c) and (d) screenshots of the reconstructed 3-D model. Red boxes in (a) and (b): feature points detected using SIFT keypoint detector. Red and blue pyramids in (c) and (d): the moving camera. The concave surfaces in the 3-D model correspond to the doors on the wall. There are two doors in (a) but three doors in (c) and (d), because only six frames are selected in (a).

truth disparity maps are not available, but visual evaluation shows that the reconstruction errors due to the lack of texture (white boxes) in Fig. 7(c) and (e) are successfully removed using our transform as shown in (d) and (f).

C. Camera Tracking

In this section, we show that our image transform can be used to estimate the camera motion for low-texture scenes. Our method differs from standard method in the first step which locates and describes the feature points in each image. Fig. 8(a) presents several frames (from the left lens of Bumblebee XB3 stereo camera) of a low-texture wall and the red boxes indicate the feature points detected using SIFT keypoint detector [26].[2] As can be seen in Fig. 8(a), no feature point is detected due to the lack of texture in the first few selected frames, it is thus impossible to estimate the camera motion from these frames. Nevertheless, applying SIFT detector to the transformed images in Fig. 8(b) shows that many feature points can be detected [red boxes in Fig. 8(b)]. In addition to the keypoint locations themselves, SIFT provides a local descriptor [computed from the transformed images in Fig. 8(b)] for each keypoint. Also, each keypoint has a depth value computed from stereo matching using the transformed images (from the left and right lens of Bumblebee XB3 stereo camera). Next, for every neighboring frame pair, we matched keypoint descriptors between them[2], and converted the matched keypoints into two 3D point clouds using the depth values at each keypoint. We finally estimate the best rotation and translation (in a Least Squares sense) that transform these two 3D point clouds using the method presented in [28]. Fig. 8(c) and (d) presents screenshots of the 3D model reconstructed using the estimated camera rotation and translation parameters, which visually demonstrate that the keypoint detection, description and the depth estimation using the transformed images are accurate. Also, from the reconstructed 3D model, we can calculate the distance between the camera centers of the first and last frame which is **15.5** meter.

[2]We use the demo program provided on the author's website to detect and match the keypoints.

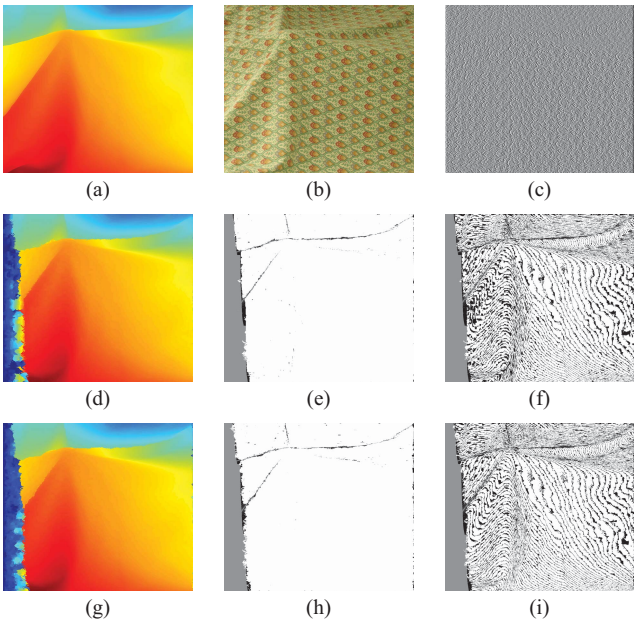


Fig. 9. Evaluation using highly textured *Cloth1* data set [27]. (a) Ground-truth disparity map and (b) and (c) reference camera image and its transform, respectively. (d) and (g) Disparity maps obtained by applying CSBP [18] to the original input stereo images and transformed images, respectively. (e) and (f) Error maps of (d) by setting the error threshold to one and half a disparity, respectively. (h) and (i) Error maps of (g) by setting error threshold to one and half a disparity, respectively. The black pixels in (e), (f), (h), and (i) are bad pixels. The gray pixels are border-occluded which are not considered in this paper. Note that the reconstruction accuracy obtained using our transform is very close to the accuracy obtained using image intensities for highly textured images captured under highly controlled environments. The percentages of bad/black pixels in (e) and (h) are 1.68% and 1.21%, respectively, and the percentages of bad/black pixels in (f) and (i) are 22.4% and 18.7%, respectively. Our transform achieves higher subpixel accuracy as it is more robust to low-texture regions. Note: The reader is urged to view these images [especially for (c)] at full size on a video display, for details may be lost in hard copy.

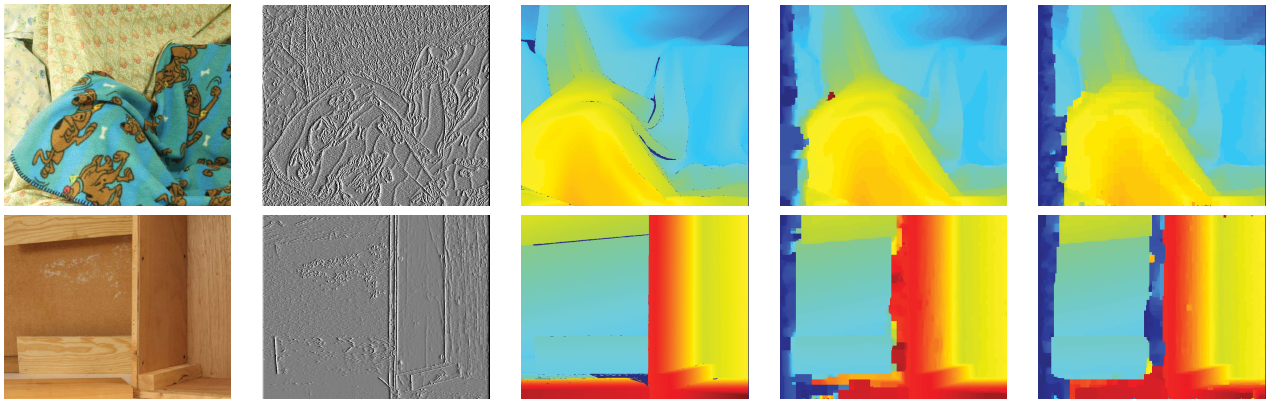


Fig. 10. Evaluation using *Cloth3* and *Cloth1* data set [27]. (a) and (b) Reference camera images and their transforms, respectively. (c) Ground-truth disparity map and (d) and (e) disparity maps obtained by applying CSBP [18] to the original input stereo images and transformed images, respectively.

For quantitative evaluation, we manually measured the distance between the positions where the first frame and the last frame were capture. The measured distance is **15.3** meter, which is close to the one estimated from images.

D. Evaluation Using Highly-Textured Middlebury Images

We have shown that our transform is more robust for matching low-texture scenes, and in this section, we will demonstrate that our transform is also robust to highly-textured scenes and can achieve similar performance as intensity-based method. Fig. 9 tested our method using the *Cloth1* data set which has the most textures among all Middlebury images [27]. The disparity error maps obtained with one disparity error are presented in Fig. 9(e) and (h), which show that the reconstruction accuracy obtained using our transform is very close to the accuracy obtained using image intensities for highly-textured images captured under highly-controlled environments. The percentages of bad/black pixels computed from (e) and (h) are **1.68%** and **1.21%**. The disparity error maps obtained with half a disparity error are presented in Fig. 9(f) and (i). The percentages of bad/black pixels computed from (f) and (i) are **22.4%** and **18.7%**. Our transform achieves higher sub-pixel accuracy as it is more robust to low-texture regions. Note that the proposed transform does not require locating of edges thus it is robust for textured images with discontinuous edges as shown in Fig. 10.

IV. Discussion

We have presented a new image transform - epipolar distance transform - in this paper. The transform captures the local image structure by computing the ratios of distances along the epipolar lines, which produce variances inside low-texture regions based on the region geometry. We theoretically prove that it is invariant to affine transformation. The transformed image can be directly used with any stereo algorithms for depth estimation. Our experiments on real images demonstrate that our transform is more reliable for matching low-texture regions, and meanwhile, robust to highly-textured regions.

One problem remaining is how to estimate the optimal value of σ_S at each pixel location. Although our experiments (Sec. III) demonstrate that the use of transformed images from constant σ_S can achieve higher reconstruction accuracy for low-texture regions, investigation into the optimal σ_S is required. Large σ_S maybe not robust to occlusions while small σ_S is invalid for large low-texture regions. Ideally, we should have large σ_S for low-texture regions and small σ_S for high-textured regions. We do not have a very neat algorithm for automatically computing the optimal σ_S at every pixel location, but a simple fusion scheme turns out to be a good solution. See Fig. 11, we compute the transformed image pair using constant σ_S according to Eqn. (5), and let them be T_L^F and T_R^F (Fig. 11(c) and (e)). The disparity map obtained from these two transformed images are presented in Fig. 11(g).

We then compute another transformed image pair using the image segmentation result according to Eqn. (1), and let them be T_L^S and T_R^S (Fig. 11(d) and (f)). In this paper, we use the real-time segmentation method presented in [29]. The segmentation method *does not need to be very stable* because only large segments will be used and unreliable segments will be detected and removed. The disparity map obtained from these two transformed images are presented in Fig. 11(h). As can be seen from Fig. 11(g) and (h), the two ways of computing the image transform are complementary: T_L^F and T_R^F are invalid for large low-textured regions (σ_S is too small to cover the whole region), while T_L^S and T_R^S are non-robust on small regions.

Our goal is fusing T_L^F and T_L^S (T_R^F and T_R^S) for a more reliable transform T_L (T_R). The basic idea for fusion is that for large and correct segments, $T_L = T_L^S$, otherwise $T_L = T_L^F$. In this paper, a correct segment on the left image mean that 99% of the pixels inside this segment can be correctly mapped to a single segment on the right image using the disparity map computed from T_L^S and T_R^S, so is its corresponding segment on the right image. We set the threshold to 99% to make sure that only very reliable segments will be used. We do not set it to 100% because we want to make sure that it is robust to noises around the edge of the segment. Also, fusion is only required for large regions as color segmentation is non-robust

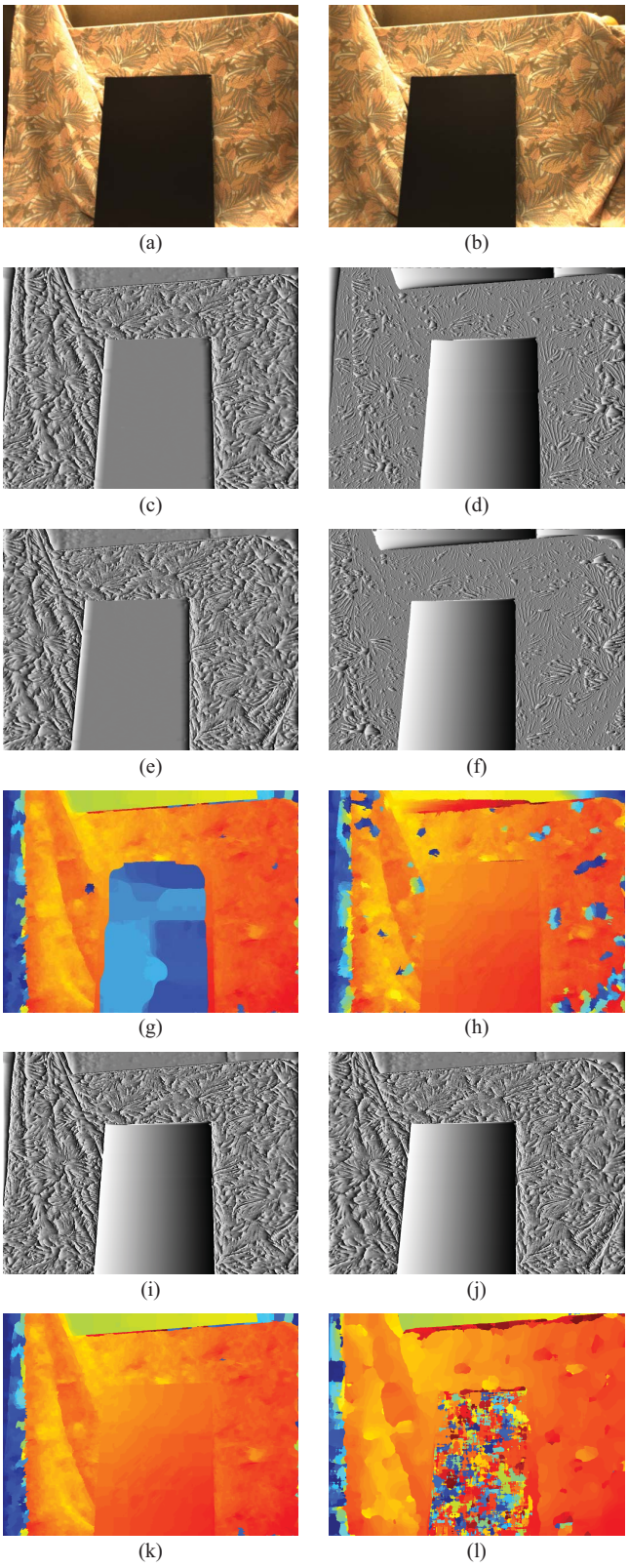


Fig. 11. (a)–(l) Optimal σ_S parameter.

for small regions. In this paper, we only consider segments at least 1% of the image size. The fused transform is presented in Fig. 11(i)–(j), and the disparity map obtained from these transform images is presented in Fig. 11(k). Visual comparison between Fig. 11(g), (h) and (k) shows that this simple fusion scheme is a good substitution when the optimal σ_S values are hard to obtain at each pixel location.

Comparing with the other transforms, the proposed method has two main advantages:

1) it works for untextured regions. Other transforms are invalid for untextured regions like the black board in Fig. 11 (a). Fig. 11 (l) is the disparity map obtained using Census transform [30]. As can be seen, Census transform fails because it measures the relative intensity values (either one or zero) which are unfortunately all zero for untextured regions.
2) its computational complexity is invariant to the disparity searching range thus is very efficient for stereo matching. Specifically, the speed of computing the transformed image of a VGA-sized RGB image ($640 \times 480 \times 3$) is about 9 frame per second on a MacBook Air laptop computer with a 1.8GHz Core i7 processor.

However, same as the other transforms, the proposed transform is invalid for occluded untextured regions.

APPENDIX
PROOF OF THEOREM 1 IN SECTION II-B

Proof: Let $\mathbf{P}$and $\mathbf{Q}$ be two endpoints of a straight line in $\mathbb{R}^3$, and $\mathbf{P}_L = [x_L^{\mathbf{P}}, y_L^{\mathbf{P}}]$, $\mathbf{Q}_L = [x_L^{\mathbf{Q}}, y_L^{\mathbf{Q}}]$ be their projections on camera C_L and $\mathbf{P}_R = [x_R^{\mathbf{P}}, y_R^{\mathbf{P}}]$, $\mathbf{Q}_R = [x_R^{\mathbf{Q}}, y_R^{\mathbf{Q}}]$ be their projections on camera C_R as shown in Fig. 1(a). Assume that the two cameras are calibrated and the captured stereo images are rectified such that the epipolar lines are scanlines, and the disparity values of pixel $\mathbf{P}_L$ and $\mathbf{Q}_L$ be $D(\mathbf{P}_L)$ and $D(\mathbf{Q}_L)$, respectively, then

$$y_R^{\mathbf{P}} = y_L^{\mathbf{P}}, \tag{7}$$
$$y_R^{\mathbf{Q}} = y_L^{\mathbf{Q}}, \tag{8}$$
$$x_R^{\mathbf{P}} = x_L^{\mathbf{P}} - D(\mathbf{P}_L), \tag{9}$$
$$x_R^{\mathbf{Q}} = x_L^{\mathbf{Q}} - D(\mathbf{Q}_L). \tag{10}$$

The angles of the straight lines $\mathbf{P}_L\mathbf{Q}_L$ and $\mathbf{P}_R\mathbf{Q}_R$ can then be represented as

$$\theta_{\mathbf{P}_L\mathbf{Q}_L} = \tan^{-1}\left(\frac{y_L^{\mathbf{Q}} - y_L^{\mathbf{P}}}{x_L^{\mathbf{Q}} - x_L^{\mathbf{P}}}\right), \tag{11}$$
$$\theta_{\mathbf{P}_R\mathbf{Q}_R} = \tan^{-1}\left(\frac{y_R^{\mathbf{Q}} - y_R^{\mathbf{P}}}{x_R^{\mathbf{Q}} - x_R^{\mathbf{P}}}\right). \tag{12}$$

Substitute Eqn. (7), (8), (9), (10) into Eqn. (12), we obtain

$$\theta_{\mathbf{P}_R\mathbf{Q}_R} = \tan^{-1}\left(\frac{y_L^{\mathbf{Q}} - y_L^{\mathbf{P}}}{x_L^{\mathbf{Q}} - x_L^{\mathbf{P}} + (D(\mathbf{P}_L) - D(\mathbf{Q}_L))}\right). \tag{13}$$

As can be seen from Eqn. (11) and (13), line $\mathbf{P}_L\mathbf{Q}_L$ has the same direction as $\mathbf{P}_R\mathbf{Q}_R$ only when

1) $\mathbf{P}_L\mathbf{Q}_L$ and $\mathbf{P}_R\mathbf{Q}_R$ are along the epipolar line: $y_L^{\mathbf{Q}} - y_L^{\mathbf{P}} = 0$;
2) or the disparity values of pixel $\mathbf{P}_L$ and $\mathbf{Q}_L$ are the same: $D(\mathbf{P}_L) - D(\mathbf{Q}_L) = 0$.

REFERENCES

[1] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 993–1008, Aug. 2003.

[2] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision*, vol. 47, nos. 1–3, pp. 7–42, Apr.–Jun. 2002.

[3] K.-J. Yoon and I.-S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 650–656, Apr. 2006.

[4] Q. Yang, "A non-local cost aggregation method for stereo matching," in *Proc. Comput. Vision Pattern Recogn.*, 2012, pp. 1–8.

[5] C. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in *Proc. ACM Trans. Graph. ACM SIGGRAPH*, 2004, pp. 600–608.

[6] W. T. Freeman, E. Pasztor, and O. T. Carmichael, "Learning low-level vision," in *Proc. IEEE Comput. Vision 17th Int. Conf.*, 1999, pp. 1182–1189.

[7] J. Sun, N. Zheng, and H. Y. Shum, "Stereo matching using belief propagation," *Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 787–800, 2003.

[8] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.

[9] D. Terzopoulos, "Regularization of inverse visual problems involving discontinuities," *Pattern Anal. Mach. Intell.*, vol. 8, no. 4, pp. 413–242, 1986.

[10] M. Bleyer and M. Gelautz, "A layered stereo algorithm using image segmentation and global visibility constraints," in *Proc. Image Process. Int. Conf.*, Oct. 2004, pp. 2997–3000.

[11] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Proc. Pattern Recogn. 18th Int. Conf.*, 2006, pp. 15–18.

[12] H. Tao, S. Harpreet, and R. Kumar, "A global matching framework for stereo computation," in *Proc. IEEE Comput. Vision 18th Int. Conf.*, 2001, pp. 532–539.

[13] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister, "Stereo matching with color-weighted correlation, hierachical belief propagation and occlusion handling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 492–504, Mar. 2009.

[14] M. Coughlan and L. Yuille, "Manhattan world: Compass direction from a single image by bayesian inference," in *Proc. IEEE Comput. Vision 17th Int. Conf.*, 1999, pp. 941–948.

[15] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski, "Manhattan-world stereo," in *Proc. IEEE Comput. Vision Pattern Recogn. Conf.*, Jun. 2009, pp. 1422–1429.

[16] S. Coorg and S. Teller, "Extracting textured vertical facades from controlled close-range imagery," in *Proc. IEEE Comput. Vision Pattern Recogn. Comput. Soc. Conf.*, 1999, pp. 625–632.

[17] T. Werner and A. Zisserman, "New techniques for automated architectural reconstruction from photographs," in *Proc. Eur. Conf. Comput. Vision*, 2002, pp. 541–555.

[18] Q. Yang, L. Wang, and N. Ahuja, "A constant-space belief propagation algorithm for stereo matching," in *Proc. IEEE Comput. Vision Pattern Recogn. Conf.*, Jun. 2010, pp. 1458–1465.

[19] N. Ahuja, "A transform for multiscale image segmentation by integrated edge and region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 12, pp. 1211–1235, Dec. 1996.

[20] S. Todorovic and N. Ahuja, "Extracting subimages of an unknown category from a set of images," in *Proc. IEEE Comput. Vision Pattern Recogn. Soc. Conf.*, Jun. 2006, pp. 927–934.

[21] S. Todorovic and N. Ahuja, "Unsupervised category modeling, recognition, and segmentation in images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2158–2174, Dec. 2008.

[22] D. Comaniciu and P. Meer, "A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[23] Q. Yang, K.-H. Tan, and N. Ahuja, "Real-time o(1) bilateral filtering," in *Proc. Comput. Vision Pattern Recogn.*, 2009, pp. 557–564.

[24] Q. Yang, S. Wang, and N. Ahuja, "Svm for edge-preserving filtering," in *Proc. IEEE Comput. Vision Pattern Recogn. Conf.*, Jun. 2010, pp. 775–1782.

[25] P. Grey. (2010). *Bumblebee xb3 Stereo Vision System* [Online]. Available: http://www.ptgrey.com/products/bbxb3/index.asp

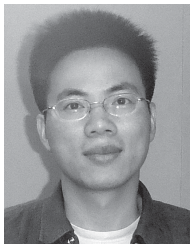
[26] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[27] D. Scharstein and R. Szeliski. (2002). *Middlebury Stereo Benchmark* [Online]. Available: http://vision.middlebury.edu/stereo/

[28] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, Sep. 1987, pp. 698–700.

[29] Q. Yang, C. Engels, and A. Akbarzadeh, "Near real-time stereo for weakly-textured scenes," in *Proc. British Mach. Conf.*, 2008, pp. 80–87.

[30] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. Eur. Conf. Comput. Vision*, 1994, pp. 151–158.



Qingxiong Yang (M'11) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana, in 2010.

He is an Assistant Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong. His current research interests include computer visions and graphics.

Dr. Yang was the recipient of the Best Student Paper Award at MMSP in 2010 and the Best Demo Award at CVPR in 2007.

Narendra Ahuja (F'92) received the B.E. degree (Hons.) in electronics engineering from the Birla Institute of Technology and Science, Pilani, India, in 1972, the M.E. degree, with distinction, in electrical communication engineering from the Indian Institute of Science, Bangalore, India, in 1974, and the Ph.D. degree in computer science from the University of Maryland, College Park, in 1979.

He was a Scientific Officer with the Department of Electronics, Government of India, New Delhi, India, from 1974 to 1975. He was with the Computer Vision Laboratory, University of Maryland, from 1975 to 1979. Since 1979, he has been with the University of Illinois at Urbana-Champaign, Urbana, where he is currently a Donald Biggar Willet Professor with the Department of Electrical and Computer Engineering, Beckman Institute, and the Coordinated Science Laboratory. He has co-authored the books *Pattern Models* (Wiley, 1983), *Motion and Structure from Image Sequences* (Springer-Verlag, 1992), and *Face and Gesture Recognition* (Kluwer, 2001); and co-edited the book *Advances in Image Understanding* (IEEE Press, 1996). His current research interests include extraction and representation of spatial structure in images and video, integrated use of multiple image-based sources for scene representation and recognition, versatile sensors for computer visions, and applications including visual communication, image manipulation, and information retrieval.

Dr. Ahuja was a recipient of the Emanuel R. Piore Award from the IEEE in 1999, the Technology Achievement Award from the International Society for Optical Engineering in 1998, the TA Stewart-Dyer/Frederick Harvey Trevithick Prize from the Institution of Mechanical Engineers in 2008, and the Open Innovation Research Award from Hewlett-Packard in 2008. He was selected as an Associate from 1998 to 1999 and from 2006 to 2007, and as a Beckman Associate from 1990 to 1991 at the University of Illinois Center for Advanced Study. He was a recipient of the Distinguished Alumnus Award from the Department of Computer Science, University of Maryland in 2008, the Best Paper Award from the IEEE Transactions on Multimedia in 2006, the University Scholar Award in 1985, the Presidential Young Investigator Award in 1984, the National Scholarship from 1967 to 1972, and the President's Merit Award in 1966. He is a fellow of the American Association for Artificial Intelligence, the International Association for Pattern Recognition, the Association for Computing Machinery, the American Association for the Advancement of Science, and the International Society for Optical Engineering. He is on the editorial boards of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Computer Vision, Graphics, and Image Processing*, the *Journal of Mathematical Imaging and Vision*, the *Journal of Pattern Analysis and Applications,* the *International Journal of Imaging Systems and Technology*, the *Journal of Information Science and Technology*, and the *IEE Japan Transactions on Electrical and Electronic Engineering*. He was a Guest Coeditor of the *Artificial Intelligence Journal*'s special issue on vision. He was the Founding Director of the International Institute of Information Technology, Hyderabad, India where he continues to serve as the Director International.